

Bioconductor CZI / HCA  
Seed Networks Symposium:  
Work in Progress

July 20, 2020

# Introduction to the symposium

Bioconductor <https://bioconductor.org>

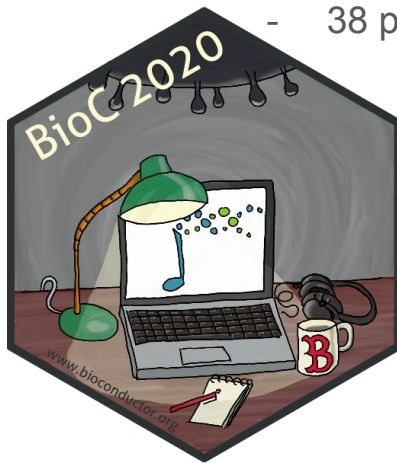
- Statistical analysis and comprehension of high-throughput genomic data
- 1900+ R packages contributed by our global user base
- Widely used & well respected

Single cell resources

- 100+ existing packages
- [Orchestrating single-cell analysis with Bioconductor](#)
- Annual conference next week!

Seed networks for the human cell atlas see

- "...bring together experimental scientists, computational biologists, software engineers, and physicians to support the continued development of the Human Cell Atlas (HCA)"
- 38 projects, of which we are one



## Bioconductor's contributions

- Access and represent HCA data
- Methods and benchmarks for emerging and integrative data
- Methods for scalable, performant analysis

## Today

- Recent updates from across our collaboration
- Short talks, with question & answer opportunities

## Project leaders

- Aedin Culhane
- Greg Finak
- Kasper Hansen
- Stephanie Hicks
- Wolfgang Huber
- Martin Morgan
- Davide Risso
- Matt Ritchie

# Data access and representation

# Programmatic access to the HCA using HCAExplorer

- Download experiment metadata and pre-computed expression matrices
- Mirrors the functionality of the HCA Data Explorer

<https://data.humancellatlas.org/explore/projects>

- <https://bioconductor.org/packages/HCAExplorerBrowser>

Human Cell Atlas DATA PORTAL Explore Guides Metadata Pipelines Analysis Tools Contribute APIs

Explore Data

Search all filters

Donor Tissue Type Specimen Method File

Genus Species: Homo sapiens Clear All

158 Donors 445 Specimens 2.7M Estimated Cells 264.3k Files 19.68 TB File Size [Export Selected Data](#)

Projects Samples Files

Project Title	Project Downloads	Species	Sample Type	Organ / Model Organ	Select	
(23)	Metadata Matrix	(1)	(3)	(17) / (6)	(32)	
<input type="checkbox"/> A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure		-	Homo sapiens, Mu...	specimens	pancreas	pancr
<input type="checkbox"/> A single-cell reference map of transcriptional states for human blood and tissue T cell activation			Homo sapiens	specimens	blood, hematopole...	T cell
<input type="checkbox"/> A single-cell transcriptome atlas of the adult human retina			Homo sapiens	specimens	eye	Unspe
<input type="checkbox"/> Assessing the relevance of organoids to model inter-individual variation			Homo sapiens	organoids	skin of body / brain	neural
<input type="checkbox"/> Bone marrow plasma cells from hip replacement surgeries		-	Homo sapiens	specimens	hematopoietic syst...	Plasm
<input type="checkbox"/> Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics		-	Homo sapiens	specimens	blood	periph
<input type="checkbox"/> Census of Immune Cells			Homo sapiens	specimens	blood, immune sys...	bone r

# Downloading an expression matrix as a LoomExperiment File using HCAExplorer

```
## Create HCAExplorer object
hca <- HCAExplorer()

## Obtain the first project by subsetting
hca <- hca[1]

## Download project's expression matrix file as a LoomExperiment object
le <- getExpressionMatrix(hca, format = "loom")
```

Select first project in the HCAExplorer object and download its matrices as a LoomExperiment object.

The screenshot shows a web browser window with the following elements:

- Browser Tab:** "Data Store API" with a close button (x) and a plus sign (+).
- Address Bar:** "data.humancellatlas.org/apis" with navigation icons (back, forward, refresh) and utility icons (star, red circle, grey circle, refresh, document, B, puzzle, profile, menu).
- Page Header:** "HUMAN CELL ATLAS DATA PORTAL" logo on the left and a "Menu" dropdown on the right.
- Main Section:** "APIs" heading followed by a link to "API Documentation".
- API List:** A list of API links: "Data Store API", "Matrix Service API", and "Data Ingest API".
- Content Area:** A large heading "Data Store API" followed by a paragraph: "The HCA Data Storage System (DSS) is a replicated data storage system designed for hosting large sets of scientific experimental data on Amazon S3 and Google Storage. The Data Store API provides a low level search and access interface to all data stored in the DCP."
- Footer:** A cookie notice: "This website uses cookies for security and analytics purposes. By using this site, you agree to these uses. Learn more [here](#)." with a "Got it" button.

# Programmatic access to the HCA DSS API using **HCABrowser**

- A more complex interface that implements the HCA DSS's api methods <https://dss.data.humancellatlas.org/>
- Mirrors the functionality of the python dcp-cli package <https://github.com/HumanCellAtlas/dcp-cli>
- <https://bioconductor.org/packages/HCABrowser>

```
(!organ.text %in% c('Brain', 'blood')) &  
(files.specimen_from_organism_json.genus_species.text == "Homo sapiens" |  
library_preparation_protocol_json.library_construction_approach.text == 'Smart-seq2'))  
...
```



Matrix Service API

data.humancellatlas.org/apis/api-documentation/ma...

HUMAN CELL ATLAS  
DATA PORTAL

Menu

# APIs

API Documentation

- Data Store API
- Matrix Service API**
- Data Ingest API

## Matrix Service API

The [Matrix Service \(MS\)](#) provides an interface to aggregate, query and access gene expression matrices stored in the [Human Cell Atlas Data Coordination Platform \(DCP\)](#).

The service exposes a [REST API](#) for querying and retrieving expression matrix

This website uses cookies for security and analytics purposes. By using this site, you agree to these uses. [Learn more here.](#)

Got it

# Programmatic access to the HCA Matrix API using `BiocManager::install("HCAMatrixBrowser")`

- Available in different formats (and representations)
  - `.loom` (LoomExperiment)
  - `.mtx` (SingleCellExperiment)
  - `.csv` (tibble list)
- Easy-to-use R interface
  - Main function: `loadHCAMatrix`
  - Primary input `bundle_fqids` -- vector of cell bundle identifiers

## Matrix Service API

The [Matrix Service](#) (MS) provides an interface to aggregate, query and access gene expression matrices stored in the [Human Cell Atlas Data Coordination Platform](#) (DCP).

The service exposes a [REST API](#) for querying and retrieving expression matrix results.

## File formats

The DCP MS enables users to prepare expression matrices in several formats by supplying the `format` parameter in the POST request to the `/matrix` endpoint. The following is a list of supported file formats:

- `.loom` (default)
- `.csv`
- `.mtx`

---

[Improve this page](#)

# HCA data on Terra

In four commands, we can obtain data from the HCA

- API endpoint:  
<https://matrix.data.humancellatlas.org/>

The screenshot displays the Terra WORKSPACES RStudio interface. At the top, the navigation bar includes 'DASHBOARD', 'DATA', 'NOTEBOOKS', 'WORKFLOWS', and 'JOB HISTORY'. A 'PLAYGROUND MODE' banner with a warning icon states: 'This feature is in early development. Your files are saved on your runtime but not to your workspace. We encourage you to save your files manually.' The RStudio menu bar includes 'File', 'Edit', 'Code', 'View', 'Plots', 'Session', 'Build', 'Debug', 'Profile', 'Tools', and 'Help'. The toolbar contains icons for file operations and a search bar. The active window is 'DESCRIPTION', showing R code for loading HCA data. The console output shows the resulting LoomExperiment object.

```
8 library(HCAMatrixBrowser)
9 hca <- HCAMatrix()
10 bundle_fqids <-
11   c("ffd3bc7b-8f3b-4f97-aa2a-78f9bac93775.2019-05-14T122736.345000Z",
12     "f69b288c-fabc-4ac8-b50c-7abcae3731bc.2019-05-14T120110.781000Z",
13     "f8ba80a9-71b1-4c15-bcfc-c05a50660898.2019-05-14T122536.545000Z",
14     "fd202a54-7085-406d-a92a-aad6dd2d3ef0.2019-05-14T121656.910000Z",
15     "fffe55c1-18ed-401b-aa9a-6f64d0b93fec.2019-05-17T233932.932000Z")
16
17 loomExp <- loadHCAMatrix(
18   api = hca, bundle_fqids = bundle_fqids, format = "loom"
19 )
```

```
~/HCAMatrixBrowser/ ↵
class: LoomExperiment
dim: 58347 5
metadata(0):
assays(1): matrix
rownames: NULL
rowData names(9): Accession Gene ... genus_species isgene
colnames(5): 3c2180aa-0aa4-411f-98dc-73ef87b447ed ceae7e4d-6871-4d47-b2af-f3c9a5b3f5db
             1cfe9423-21d1-4281-9f9d-3aaa07b8e1e8 a2a2f604-444c-41b1-befa-25cf7461bf74
             1c2e0012-28f1-4466-92c7-d11ba756c89b
colData names(38): CellID barcode ... specimen_from_organism.provenance.document_id total_umis
rowGraphs(0): NULL
colGraphs(0): NULL
> |
```

# rhdf5 can read files in S3 buckets

- Latest version of rhdf5 distributed with support for reading directly from S3 e.g.

```
public_S3_url <-  
'https://rhdf5-public.s3.eu-central-1.amazonaws.com/rhdf5ex_t_float_3d.h5'
```

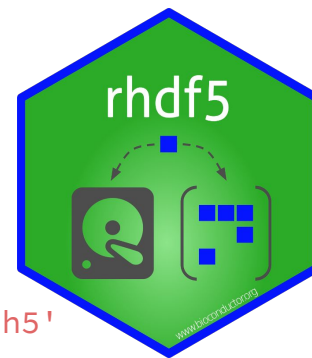
## ## EXPLORE FILE ##

```
h5ls(file = public_S3_url, s3 = TRUE)  
##   group name      otype dclass      dim  
## 0    /    a1 H5I_DATASET  FLOAT 5 x 10 x 2
```

## ## READ SUBSET ##

```
h5read(public_S3_url, name = "a1", index = list(1:2, 3, 1), s3 = TRUE)  
##   , , 1  
##           [,1]  
## [1,] 0.2444485  
## [2,] 0.3873723
```

- Works with public and private buckets



# rhdf5filters provides additional compression filters in R



- Currently seven filters:
  - BLOSC meta compressor (6 filters)
  - BZIP2
- Compiles C code on all platforms (inc Windows) no pre-built binary required
- Integrated with **rhdf5**
  - Writing: Supply filter argument to functions
  - Reading: Used automatically if needed
- <https://bioconductor.org/packages/rhdf5filters/>
- Future plan: integrate all plugins distributed by HDF5 Group ([link](#))